**Capstone Project** **– Modelling Motorbike Metrics**

Institute of Data – Data Science and AI Course

Part-time – 23 June 2020 Cohort

# Ciaran Mathewson

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.

Diagram

Description automatically generated

# Problem statement

* This project investigates the features behind what causes motorcycle crashes within Australia. It also aims to determine the relationship between increases in motorbike registrations and crash statistics.
* Public benefits
  + Lower purchasing, rego, repairs, insurance costs (than cars)
  + Less for tolls and free parking
  + Efficient commuting
  + Rapid social interaction
* Business benefits
  + Less damage to infrastructure, cheaper road upkeep
  + Easier traffic management, breakdowns do not cause blockages,
  + Lower insurance payouts
  + Environmentally ‘friendly’, less space for parking so less need to concrete large areas
* Currently, crashes are increasing as are motorbike registrations. We do not want these to be increasing at the same rate.

# Industry/ domain

* This project is within the government industry, relating to traffic conditions and crash statistics
* This project also has links to retail, as a positive result could boost sales.

# Stakeholders

* The stakeholders of this project are consumers (including myself) who purchase motorcycles, the retailers of motorcycles, insurance companies, infrastructure companies, traffic analysts and environmental scientists
* Consumers, infrastructure and insurance companies want to pay less money, whereas motorcycle retailers want to make more. As motorcycles have excellent traffic flow, traffic analysts need less rigorous measures to adapt traffic strategies for motorcycles. Motorbikes have less emissions as well as use less petrol, so benefit in reducing climate change and contribute less to the oil industry.

# Business question

* The main business question is this:
  + Can the most likely motorcycle crash scenarios be predicted?
* This would target solutions for lowering the chance of crashes, increasing confidence in new and existing riders and thus flow on to increase bike sales.
* Better than random chance (50%) would be the minimum accuracy for a model, ideally over 75% accuracy.

# Data question

* 1: What are the most important features in motorbike crashes?
* 2: Has increased motorbike usage increased injury rate?

# Data

* The dataset is the Australia & New Zealand Road Crash Dataset, accessed for free on Kaggle: <https://www.kaggle.com/mgray39/australia-new-zealand-road-crash-dataset>
* The dataset contains over 1.1 million rows, separated into six tables.
* The dataset has been collated and collaborated from a mixture of different states’ crash statistics data, so there are multiple missing values.

# Data science process

## Data analysis

* The raw data needed to be converted into individual tables, and these tables needed to be joined and all missing values either removed or imputed.
* What are the highlights of the Exploratory Data Analysis (EDA)?
  + Crash chances are increased depending on the time of day, the weekday and the speed limit of the road:

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

## Modelling

* What are the main features used?
  + 27 features at the start, reduced to 8 main features by the end
* Did you find any interesting interactions between features?
  + There was a higher chance of crashes at a lower speed, rather than a higher!
* Is there a subset of features that would get a significant portion of your final performance? Which features?
  + Hour, Month, Weekday and Speed Limit
* How did you select features?
  + Random Forest Feature Importance
* What are the models used?
  + Tried Random Forest, Logistic Regression, Decision Trees, Ada- and XG-Boost
* What are the tools used? (cloud platform, for example)
  + Jupyter Notebook through Anaconda Navigator
* What are the model performance metrics?
  + The most important was recall, but precision and overall accuracy were important (classification task). 73% recall and 71% accuracy were the best.
* Which model was selected?
  + A random forest with optimised hyperparameters was the best model

## Outcomes

* What are the main findings and conclusions of the data science process?
  + A model can be created, however the sheer amount of features makes fitting the data difficult, as there is a significant amount of noise.

# Data answer

* Was the data question answered satisfactorily?
  + Both data questions were answered, with a high level of confidence, mainly from the EDA before and after the modelling process rather than the model itself.

# Business answer

* Was the business question answered satisfactorily?
  + It was answered, but without a perfect degree of confidence (the best accuracy was less than 75%)

# 

# Response to stakeholders

* What are the overall message and recommendations to the stakeholders?
  + There are many features that influence motorcycle crashes, mainly time and date related. Variable speed limits could be implemented in these time frames to lower the risk of crashes. As commuting times were high-risk, increased awareness campaigns could prevent accidental car-motorcycle incidents.

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?
  + The model could be used mainly as a feature identification tool, rather than a predictive measure.

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)
  + Zipped in my GitHub (<https://github.com/ciaranphilip/cpm-iod>)
  + Jupyter Notebooks included in GitHub
  + Dataset accessed here: <https://www.kaggle.com/mgray39/australia-new-zealand-road-crash-dataset>
  + Images from google image search
  + Motorcycle statistics and inspiration from:
    - Christie R. & Newland, R. 2006, “Motorcyclist Fatality and Motorcycle Sales Patterns in Australia: An Update”, Accessed at: <http://casr.adelaide.edu.au/rsr/RSR2006/ChristieR.pdf>
    - Federal Chamber of Automotive Industries – Motorcycling in Australia, Directions for the Motorcycle Industry 2014-2016, Accessed at: <https://www.fcai.com.au/library/publication/Motorcycling%20in%20Australia%20-%20Directions%20for%20the%20Motorcycle%20Industry%202014-2016.pdf>
* What are the resources used in the project? (libraries, algorithms, etc)
  + Numpy, sklearn, pandas, matplotlib, seaborn, imblearn (SMOTE)
  + All included in my Jupyter Notebook